Sherlock Rules

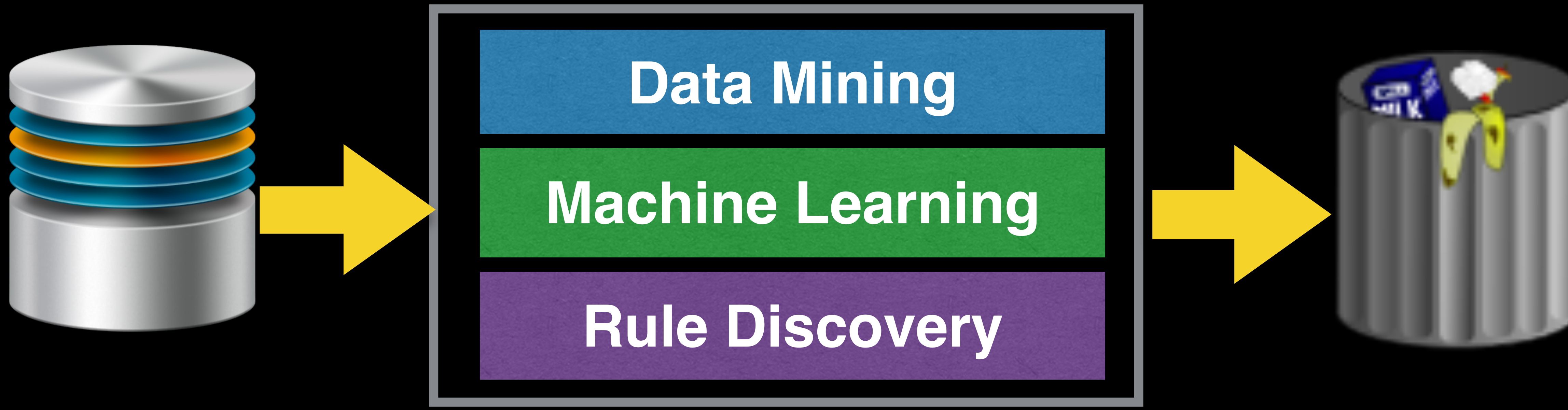
Proof Positive and Negative in Data Cleaning

Matteo Interlandi
Nan Tang

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

Outline

- **Motivation**
- Sherlock Rules
- Fundamental problems
- Algorithms

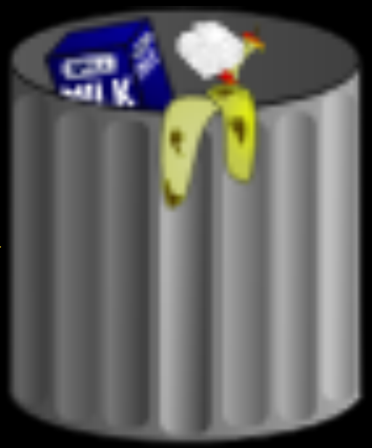


Roadblocks to Get Value from Data?

Data Mining

Machine Learning

Rule Discovery



Roadblocks to Get Value from Data?

$3 Trillion Problem: Three Best P
Today's Dirty Data Pandemic

Maybe your software is healthy, but is your data terminally ill?

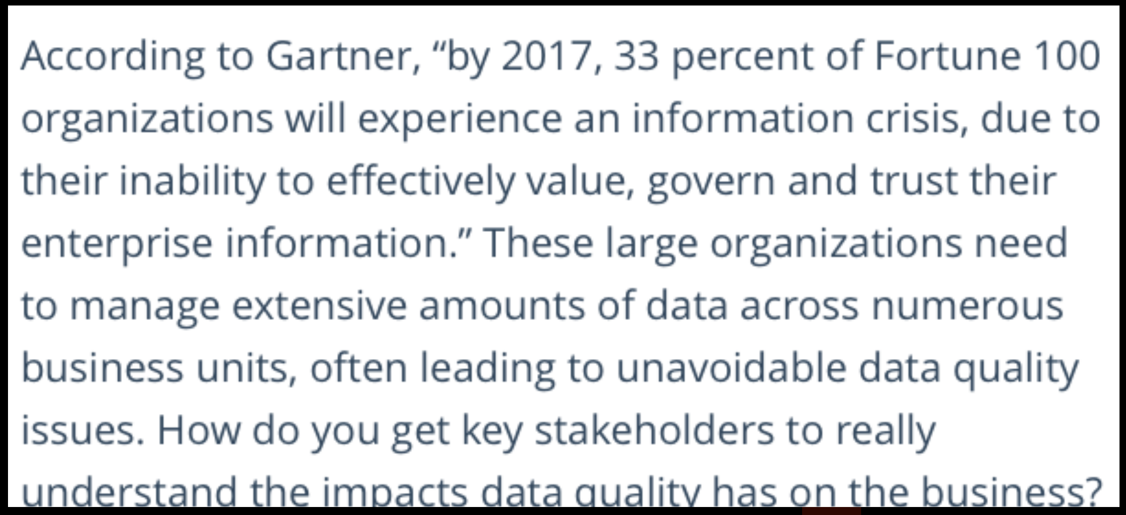
BY **HOLLIS TIBBETTS** ARTICLE RATING:

SEPTEMBER 10, 2011 12:00 PM EDT READS: 17,513

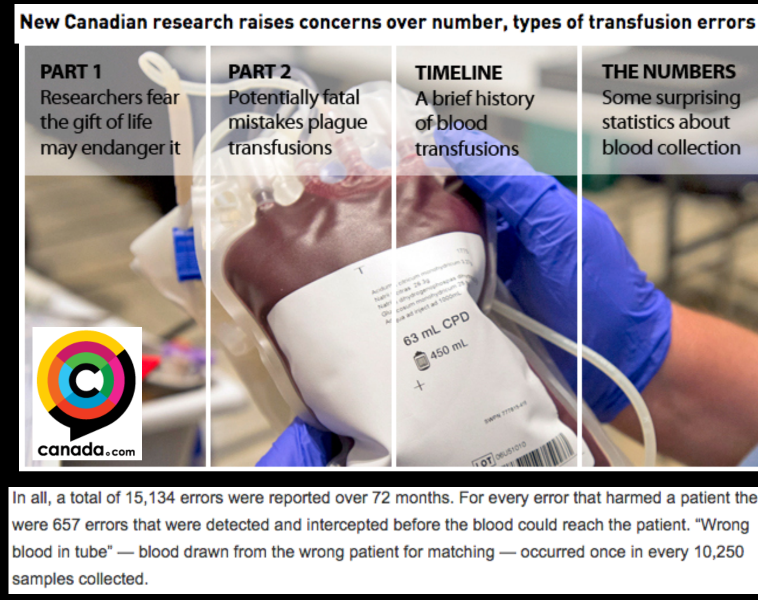
RELATED PRINT EMAIL FEEDBACK ADD THIS BLOG THIS

In survey after survey, about half of IT executives consistently agree that data quality and data consistency is one of the biggest roadblocks to them getting full value from their data.

This has been consistently true all since the Chinese invented the abacus. I suspect it will be true long after quantum computing has solved every other problem that humanity faces.



According to Gartner, "by 2017, 33 percent of Fortune 100 organizations will experience an information crisis, due to their inability to effectively value, govern and trust their enterprise information." These large organizations need to manage extensive amounts of data across numerous business units, often leading to unavoidable data quality issues. How do you get key stakeholders to really understand the impacts data quality has on the business?



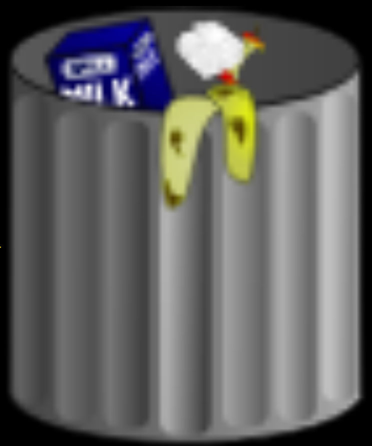
New Canadian research raises concerns over number, types of transfusion errors

In all, a total of 15,134 errors were reported over 72 months. For every error that harmed a patient there were 657 errors that were detected and intercepted before the blood could reach the patient. "Wrong blood in tube" — blood drawn from the wrong patient for matching — occurred once in every 10,250 samples collected.

Data Mining

Machine Learning

Rule Discovery



Roadblocks to Get Value from Data?

High Quality Data

$3 Trillion Problem: Three Best P… Today's Dirty Data Pandemic

Maybe your software is healthy, but is your data terminally ill?

BY HOLLIS TIBBETTS — ARTICLE RATING:

SEPTEMBER 10, 2011 12:00 PM EDT — READS: 17,513

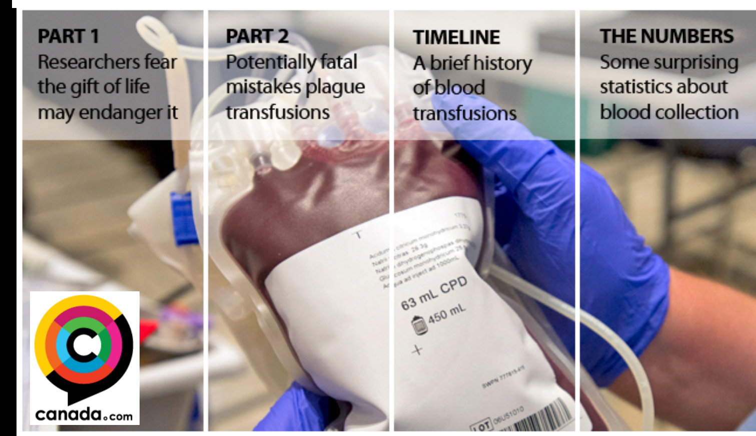
RELATED · PRINT · EMAIL · FEEDBACK · ADD THIS · BLOG THIS

In survey after survey, about half of IT executives consistently agree that data quality and data consistency is one of the biggest roadblocks to them getting full value from their data.

This has been consistently true all since the Chinese invented the abacus. I suspect it will be true long after quantum computing has solved every other problem that humanity faces.

According to Gartner, "by 2017, 33 percent of Fortune 100 organizations will experience an information crisis, due to their inability to effectively value, govern and trust their enterprise information." These large organizations need to manage extensive amounts of data across numerous business units, often leading to unavoidable data quality issues. How do you get key stakeholders to really understand the impacts data quality has on the business?

New Canadian research raises concerns over number, types of transfusion errors



In all, a total of 15,134 errors were reported over 72 months. For every error that harmed a patient the were 657 errors that were detected and intercepted before the blood could reach the patient. "Wrong blood in tube" — blood drawn from the wrong patient for matching — occurred once in every 10,250 samples collected.

D

name	nation	capital
Si	China	Beijing
Yan	China	Shanghai
Ian	China	Tokyo

D

name	nation	capital
Si	China	Beijing
Yan	China	Shanghai
Ian	China	Tokyo

data repairing

consistent D'

nation -> capital

name	nation	capital
Si	China	Beijing
Yan	China	**Beijing**
Ian	China	**Beijing**

D

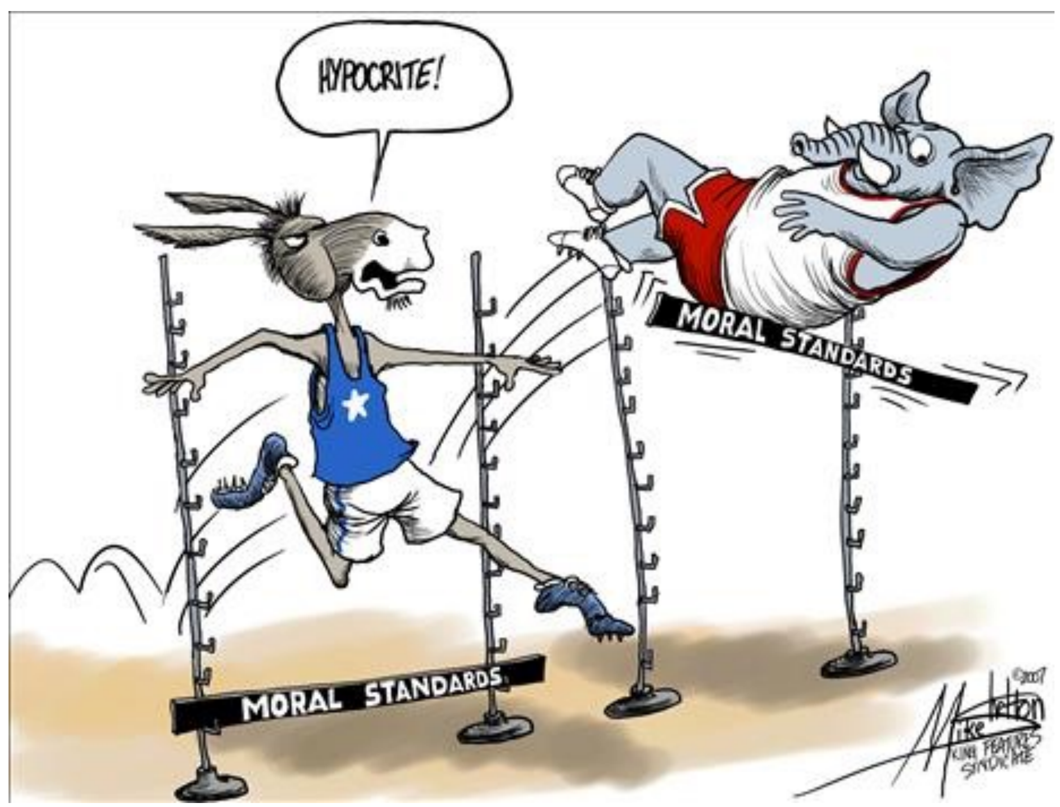
name	nation	capital
Si	China	Beijing
Yan	China	Shanghai
Ian	China	Tokyo

data repairing

consistent D'

nation -> capital

name	nation	capital
Si	China	Beijing
Yan	China	**Beijing**
Ian	China	**Beijing**



D

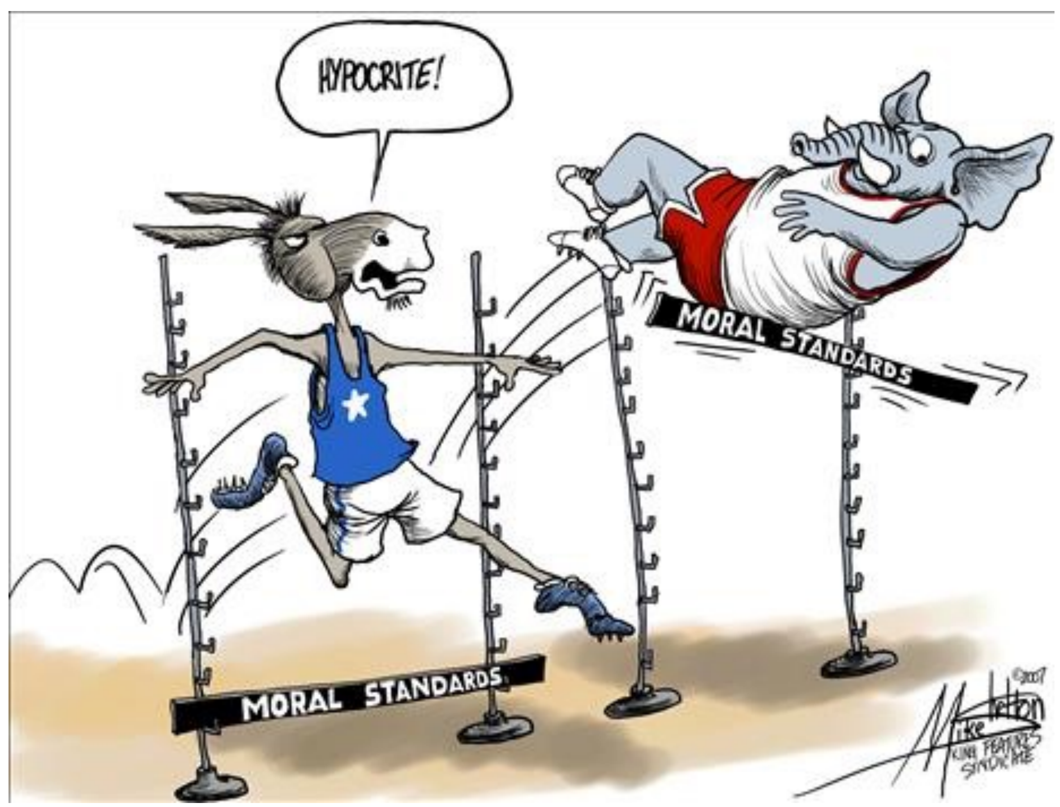
name	nation	capital
Si	China	Beijing
Yan	China	Shanghai
Ian	China	Tokyo

proof positive and negative

data repairing

annotated D''

name	nation	capital
Si	China	Beijing
Yan	China	Shanghai
Ian	China	Tokyo



consistent D'

nation -> capital

name	nation	capital
Si	China	Beijing
Yan	China	**Beijing**
Ian	China	**Beijing**

D

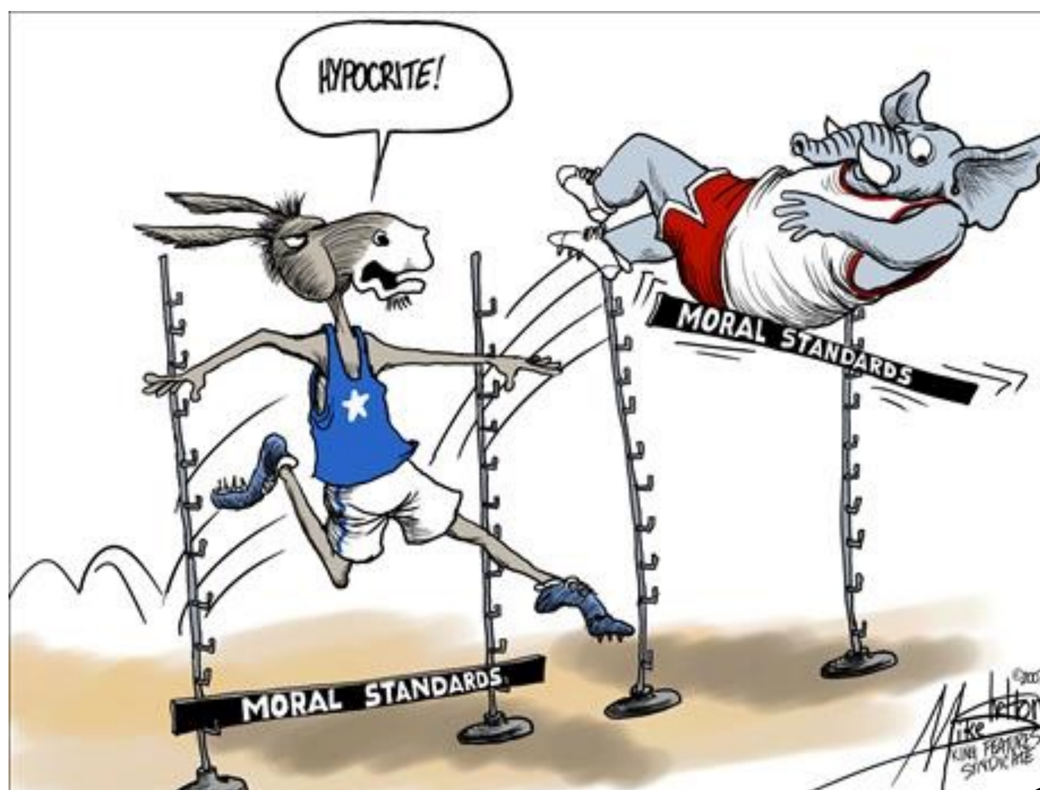
proof positive and negative

data repairing

name	nation	capital
Si	China	Beijing
Yan	China	Shanghai
Ian	China	Tokyo

annotated D''

name	nation	capital
Si	China	Beijing
Yan	China	Shanghai
Ian	China	Tokyo



consistent D'

nation -> capital

name	nation	capital
Si	China	Beijing
Yan	China	**Beijing**
Ian	China	**Beijing**

help

D

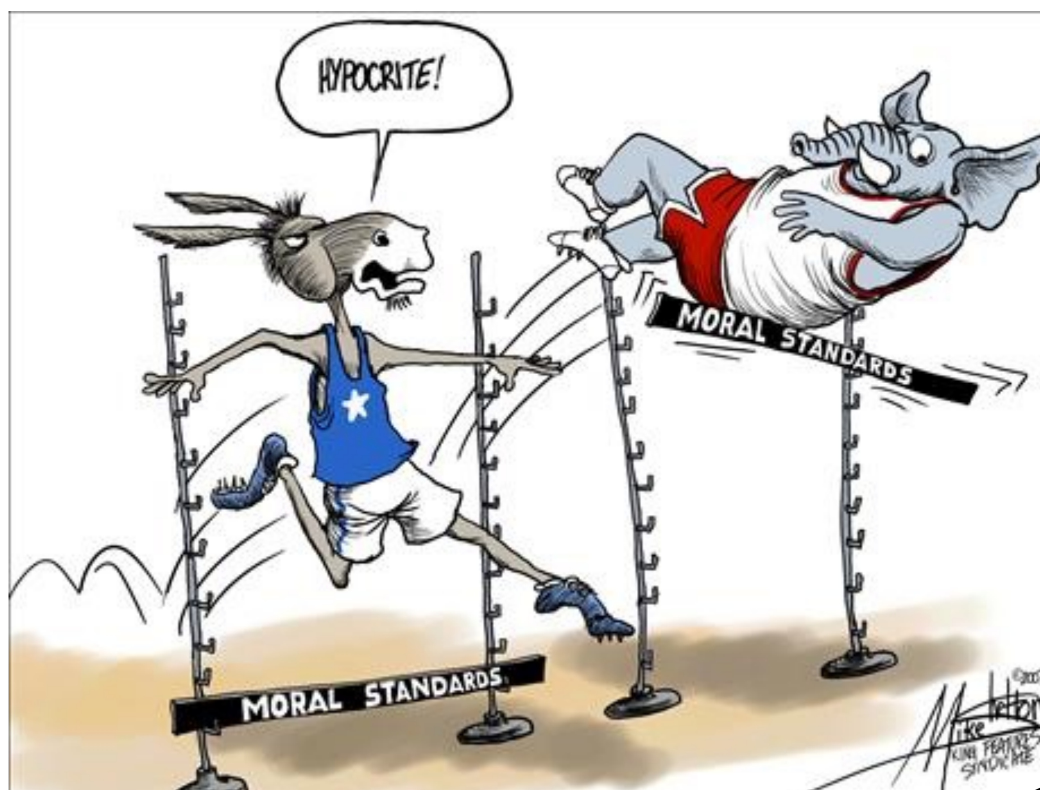
proof positive
and negative

name	nation	capital
Si	China	Beijing
Yan	China	Shanghai
Ian	China	Tokyo

data repairing

annotated D''

name	nation	capital
Si	China	Beijing
Yan	China	Shanghai
Ian	China	Tokyo



consistent D'

nation -> capital

name	nation	capital
Si	China	Beijing
Yan	China	**Beijing**
Ian	China	**Beijing**

Sherlock Rules

help

Outline

- Motivation
- **Sherlock Rules**
- Fundamental problems
- Algorithms

Proof Positive and Negative

	name	dep	nation	capital	bornat	officePhn
t1	Si	DA	China	Beijing	ChenYang	28098001
t2	Yan	DA	China	Shanghai	Chengdu	24038698
t3	Ian	ALT	China	Beijing	Hangzhou	33668323

	name	officePhn	mobile
r1	Si	28098001	66700541
r2	Yan	24038698	66706563
r3	Ian	27364928	33668323

Proof Positive and Negative

	name	dep	nation	capital	bornat	officePhn
t1	Si	DA	China	Beijing	ChenYang	28098001
t2	Yan	DA	China	Shanghai	Chengdu	24038698
t3	Ian	ALT	China	Beijing	Hangzhou	33668323

	name	officePhn	mobile
r1	Si	28098001	66700541
r2	Yan	24038698	66706563
r3	Ian	27364928	33668323

Proof Positive and Negative

	name	dep	nation	capital	bornat	officePhn
t1	Si	DA	China	Beijing	ChenYang	28098001
t2	Yan	DA	China	Shanghai	Chengdu	24038698
t3	Ian	ALT	China	Beijing	Hangzhou	33668323

	name	officePhn	mobile
r1	Si	28098001	66700541
r2	Yan	24038698	66706563
r3	Ian	27364928	33668323

Proof Positive/Negative, Correction

t3[Ian] is correct,
t3[officePhn] = 27364928

Proof Positive and Negative

	name	dep	nation	capital	bornat	officePhn
t1	Si	DA	China	Beijing	ChenYang	28098001
t2	Yan	DA	China	Shanghai	Chengdu	24038698
t3	Ian	ALT	China	Beijing	Hangzhou	33668323

	name		mobile
r1	Si		66700541
r2	Yan		66706563
r3	Ian		33668323

Proof Positive/Negative, Correction

t3[Ian] is correct,
t3[officePhn] = 27364928

Proof Positive and Negative

	name	dep	nation	capital	bornat	officePhn
t1	Si	DA	China	Beijing	ChenYang	28098001
t2	Yan	DA	China	Shanghai	Chengdu	24038698
t3	Ian	ALT	China	Beijing	Hangzhou	33668323

	name		mobile
r1	Si		66700541
r2	Yan		66706563
r3	Ian		33668323

Proof Positive/Negative, Correction

t3[Ian] is correct,
t3[officePhn] = 27364928

Proof Positive/Negative

t3[Ian] is correct,
t3[officePhn] is wrong

Proof Positive and Negative

	name	dep	nation	capital	bornat	officePhn
t1	Si	DA	China	Beijing	ChenYang	28098001
t2	Yan	DA	China	Shanghai	Chengdu	24038698
t3	Ian	ALT	China	Beijing	Hangzhou	33668323

	country	capital
s1	China	Beijing
s2	Japan	Tokyo
s3	Chile	Santiago

Proof Positive/Negative, Correction

t3[Ian] is correct,
t3[officePhn] = 27364928

Proof Positive/Negative

t3[Ian] is correct,
t3[officePhn] is wrong

Proof Positive and Negative

	name	dep	nation	capital	bornat	officePhn
t1	Si	DA	China	Beijing	ChenYang	28098001
t2	Yan	DA	China	Shanghai	Chengdu	24038698
t3	Ian	ALT	China	Beijing	Hangzhou	33668323

	country	capital
s1	China	Beijing
s2	Japan	Tokyo
s3	Chile	Santiago

Proof Positive/Negative, Correction

t3[Ian] is correct,
t3[officePhn] = 27364928

Proof Positive/Negative

t3[Ian] is correct,
t3[officePhn] is wrong

Proof Positive

t1[nation, capital] is correct
t3[nation, capital] is correct

Sherlock Rules

D

	name	dep	nation	capital	bornat	officePhn
t1	Si	DA	China	Beijing	ChenYang	28098001
t2	Yan	DA	China	Shanghai	Chengdu	24038698
t3	Ian	ALT	China	Beijing	Hangzhou	33668323

D_m

	name	officePhn	mobile
r1	Si	28098001	66700541
r2	Yan	24038698	66706563
r3	Ian	27364928	33668323

	country	capital
s1	China	Beijing
s2	Japan	Tokyo
s3	Chile	Santiago

evidence positive

$$\varphi : ((X, X_m), (B, B_m^-, B_m^+), \vec{\approx})$$

negative

Sherlock Rules

D

	name	dep	nation	capital	bornat	officePhn
t1	Si	DA	China	Beijing	ChenYang	28098001
t2	Yan	DA	China	Shanghai	Chengdu	24038698
t3	Ian	ALT	China	Beijing	Hangzhou	33668323

Dm

	name	officePhn	mobile
r1	Si	28098001	66700541
r2	Yan	24038698	66706563
r3	Ian	27364928	33668323

	country	capital
s1	China	Beijing
s2	Japan	Tokyo
s3	Chile	Santiago

evidence

positive

$$\varphi : ((X, X_m), (B, B_m^-, B_m^+), \vec{\approx})$$

negative

φ_1: ((name, name), (officePhn, mobile, officePhn), $(=, =, =)$)

φ_2: ((name, name), (officePhn, mobile, $\perp$), $(=, =, \not\approx)$)

φ_3: ((nation, country), (capital, $\perp$, capital), $(=, \not\approx, =)$)

Point of Innovation

Integrity Constraints

There does not exist

$t1[X1] = t2[X2]$ but

$t1[B1] = t2[B2]$

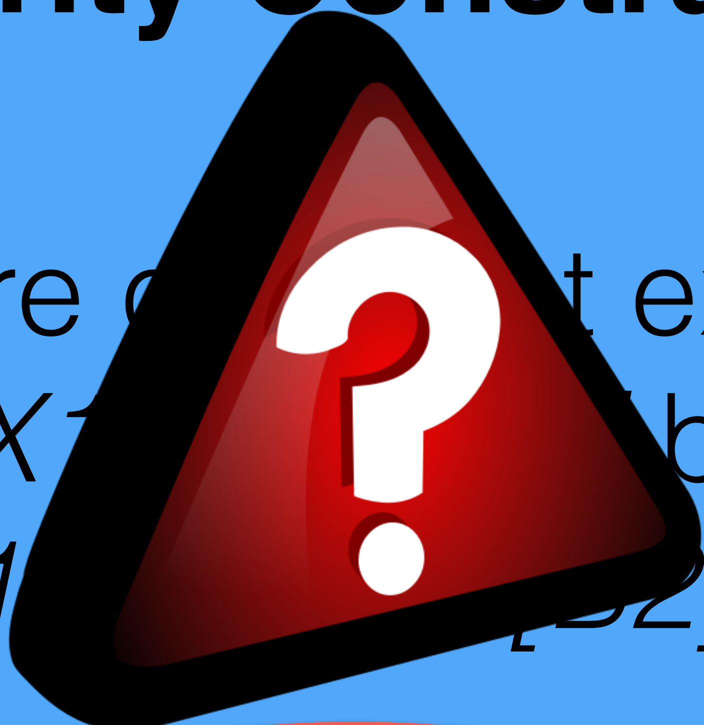
(China, Shanghai)

|| <>

(China, Beijing)

Point of Innovation

Integrity Constraints

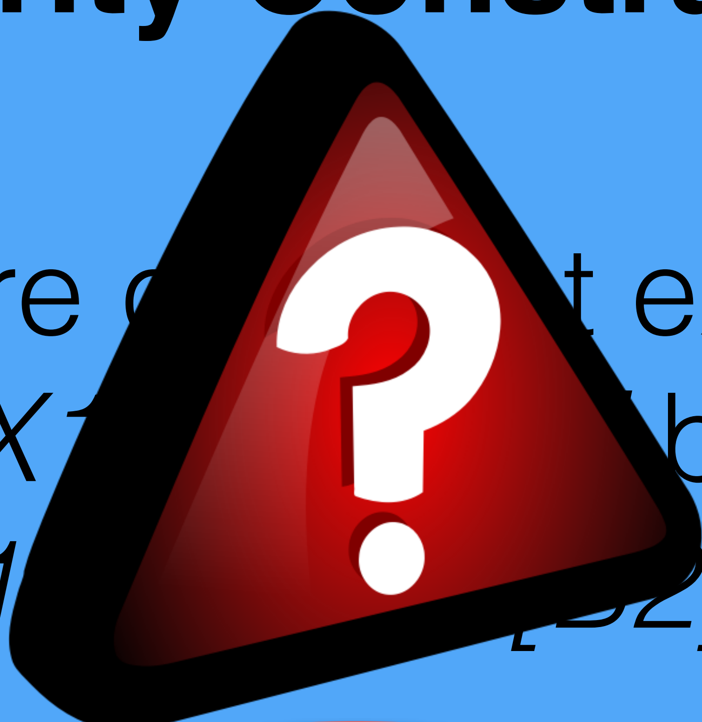


There d… t exist
t1[X1 … *] but*
t1 … *[B2]*

(China, Shanghai)
|| ∧∨
(China, Beijing)

Point of Innovation

Integrity Constraints	**Sherlock Rules**
There … t exist	*t1[X1] = t2[X2]* and
t1[X1 …] but	*t1[B] = t2[B⁻], then*
t1 … [B2]	*t1[B] := t2[B⁺]*
(China, Shanghai)	(China, Shanghai)
|| ^ v	
(China, Beijing)	(China, Beijing, Shanghai)



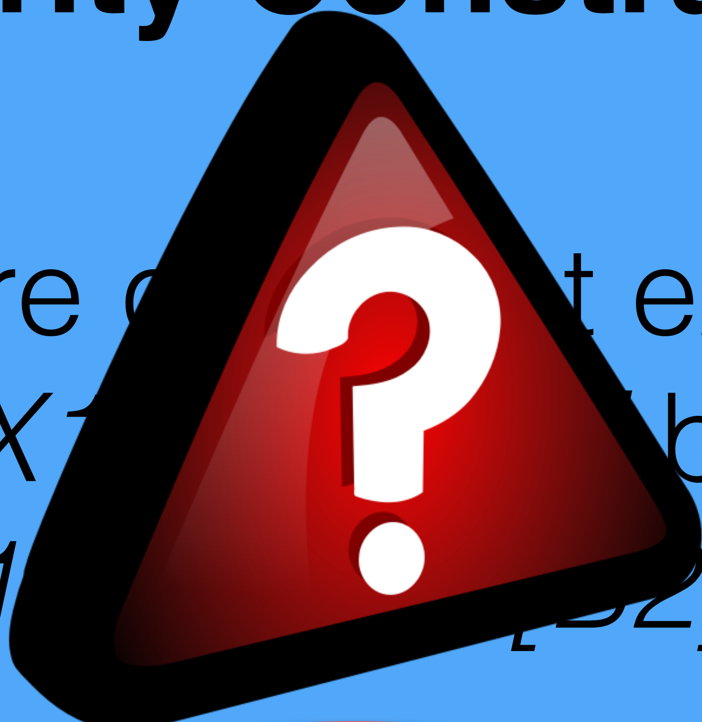
Point of Innovation

Integrity Constraints

There c… t exist

t1[X1 … but

t1 … *[B2]*



(China, Shanghai)

II ∧∨

(China, Beijing)

Sherlock Rules

$t1[X1] = t2[X2]$ and

$t1[B] = t2[B^-]$, *then*

$t1[B] := t2[B^+]$

(China, Shanghai)

(China, Beijing, Shanghai)

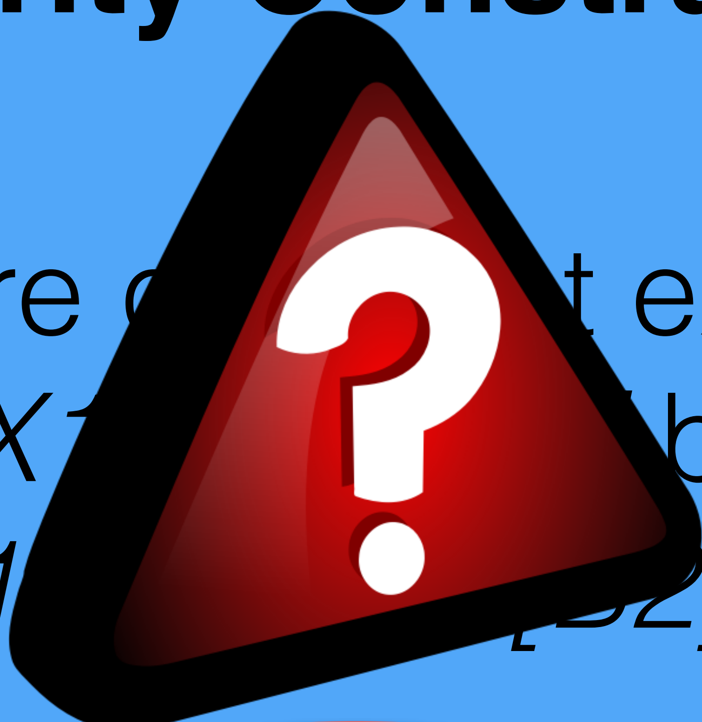
Point of Innovation

Integrity Constraints

There c… t exist

t1[X1 … but

t1 … *[B2]*



(China, Shanghai)

= <>

(China, Beijing)

Sherlock Rules

$t1[X1] = t2[X2]$ and

$t1[B] = t2[B^-]$, *then*

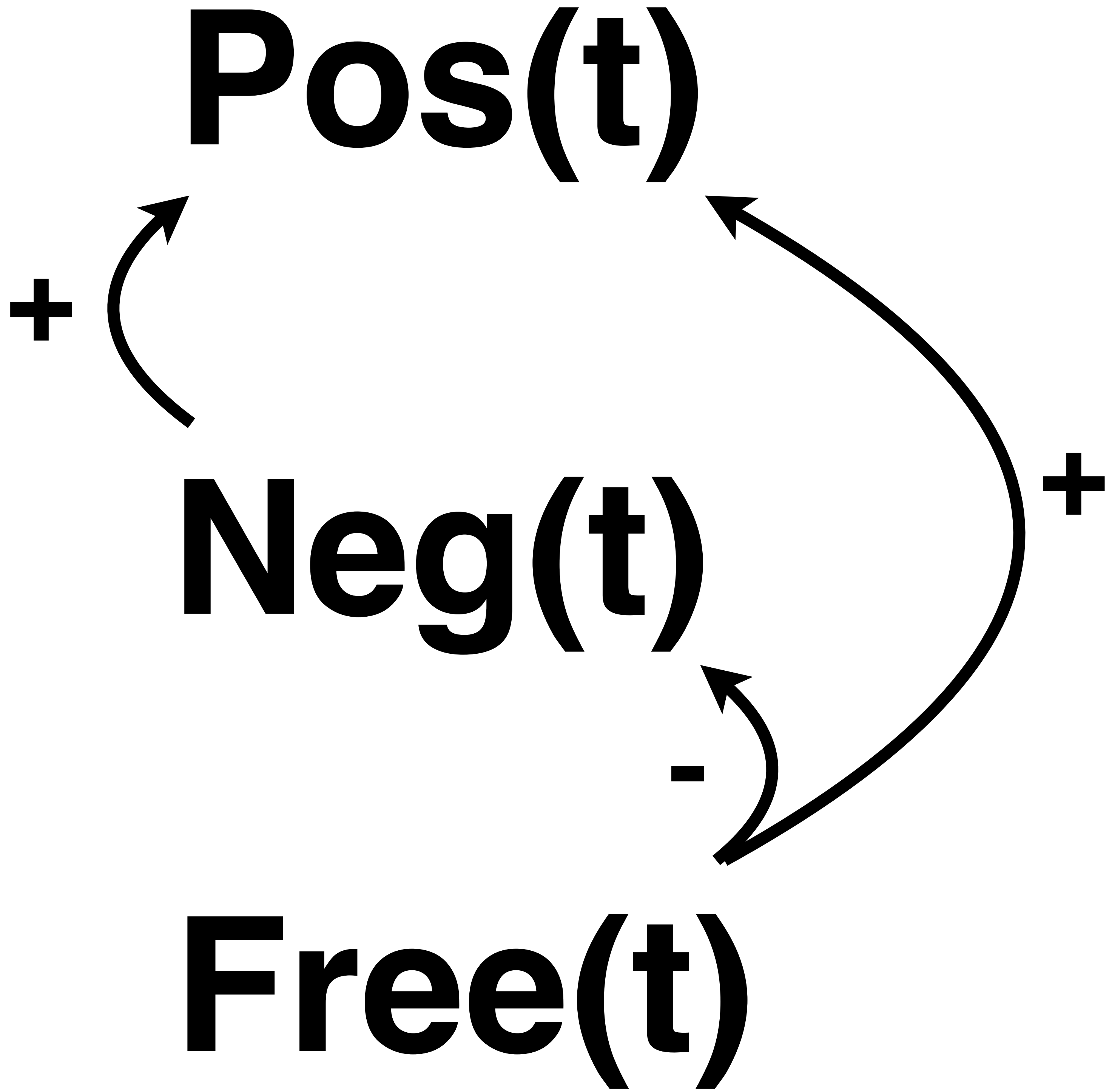
$t1[B] := t2[B^+]$

(China, Shanghai)

\+ \+ −

(China, Beijing, Shanghai)

Applying Multiple Rules



Sherlock Rules in Action

t1 (Si, DA, China, Beijing, ChenYang, 28098001)

↓

*t1 (**Si**+, DA, China, Beijing, ChenYang-, **28098001**+)*

↓

*t1 (**Si**+, DA, China, Beijing, **ShenYang**+, **28098001**+)*

Sherlock Rules in Action

t1 (Si, DA, China, Beijing, ChenYang, 28098001)

*t1 (**Si+**, DA, China, Beijing, ChenYang-, **28098001+**)*

*t1 (**Si+**, DA, China, Beijing, **ShenYang+**, **28098001+**)*

Pos(t1)

Transformation Rules

$$\frac{(X_m \neq \bot) \wedge (B_m^- \neq \bot) \wedge (B_m^+ \neq \bot) \wedge (B \notin \textsc{pos}(t)) \wedge (X \cap \textsc{neg}(t) = \bot) \wedge (t[X] \approx t_m[X_m]) \wedge (t[B] \approx t_m[B_m^-])}{(t[X, B] := t_m[X_m, B_m^+]) \wedge (\textsc{pos}(t) := \textsc{pos}(t) \cup X \cup \{B\}) \wedge (\textsc{neg}(t) := \textsc{neg}(t) \setminus \{B\})} (1)$$

$$\frac{(X_m \neq \bot) \wedge (B_m^- \neq \bot) \wedge (B_m^+ = \bot) \wedge (B \notin \textsc{pos}(t)) \wedge (X \cap \textsc{neg}(t) = \bot) \wedge (t[X] \approx t_m[X_m]) \wedge (t[B] \approx t_m[B_m^-])}{(t[X] := t_m[X_m]) \wedge (\textsc{pos}(t) := \textsc{pos}(t) \cup X) \wedge (\textsc{neg}(t) := \textsc{neg}(t) \cup \{B\})} (2)$$

$$\frac{(X_m \neq \bot) \wedge (B_m^+ \neq \bot) \wedge (B_m^- = \bot) \wedge (B \notin \textsc{pos}(t)) \wedge (B \notin \textsc{neg}(t)) \wedge (X \cap \textsc{neg}(t) = \bot) \wedge (t[X] \approx t_m[X_m]) \wedge (t[B] \approx t_m[B_m^+])}{(t[X, B] := t_m[X_m, B_m^+]) \wedge (\textsc{pos}(t) := \textsc{pos}(t) \cup X \cup \{B\})} (3)$$

$$\frac{(X_m \neq \bot) \wedge (B_m^+ \neq \bot) \wedge (B_m^- = \bot) \wedge (B \notin \textsc{pos}(t)) \wedge (X \subseteq \textsc{pos}(t)) \wedge (t[X] \approx t_m[X_m]) \wedge (t[B] \not\approx t_m[B_m^+])}{(t[B] := t_m[B_m^+]) \wedge (\textsc{pos}(t) := \textsc{pos}(t) \cup \{B\}) \wedge (\textsc{neg}(t) := \textsc{neg}(t) \setminus \{B\})} (4)$$

$$\frac{(X_m = \bot) \wedge (B_m^+ \neq \bot) \wedge (B_m^- \neq \bot) \wedge (B \notin \textsc{pos}(t)) \wedge (t[B] \approx t_m[B_m^-])}{(t[B] := t_m[B_m^+]) \wedge (\textsc{pos}(t) := \textsc{pos}(t) \cup \{B\})} (5)$$

Outline

- Motivation
- Sherlock Rules
- **Fundamental problems**
- Algorithms

Fundamental Problems

Termination	**Consistency** (*coNP-complete*)
Determinism	**Implication** (*coNP-complete*)

Algorithms

- Motivation
- Sherlock Rules
- Fundamental problems
- **Algorithms**

Algorithms

Naive Repairing

chase-based

O(|R|x|Sigma|x|M|)

Algorithms

Naive Repairing

chase-based

O(|R|x|Sigma|x|M|)

Fast Repairing

Similarity indices
to reduce |M|
(BK-tree, FastSS, n-gram)

Inverted index
to reduce |Sigma|
(hash map)

O(|R|x|Sigma| x com(S))

Algorithms

Naive Repairing

chase-based

O(|R|x|Sigma|x|M|)

Fast Repairing

Similarity indices
to reduce |M|
(BK-tree, FastSS, n-gram)

Inverted index
to reduce |Sigma|
(hash map)

Caching similarity index accesses
Rule pruning based on dependency

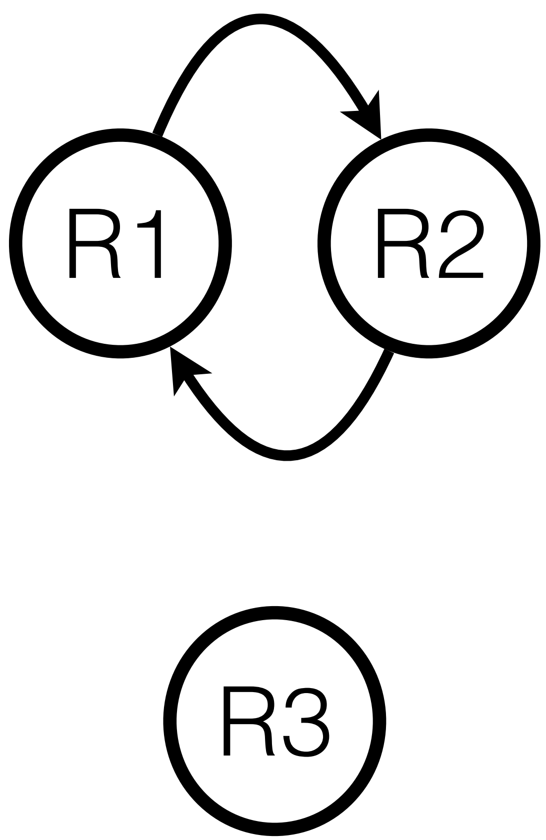
Rule Pruning Example

R1: $((\text{name}, \text{name}), (\text{officePhn}, \text{mobile}, \text{officePhn}), (=, =, =))$

R2: $((\text{name}, \text{name}), (\text{bornat}, \perp, \text{borncity}), (=, \not\approx, =))$

R3: $((\text{nation}, \text{country}), (\text{capital}, \perp, \text{capital}), (=, \not\approx, =))$

t3(Ian, ALT, Chine, Beijing, Hangzhou, 33668323)



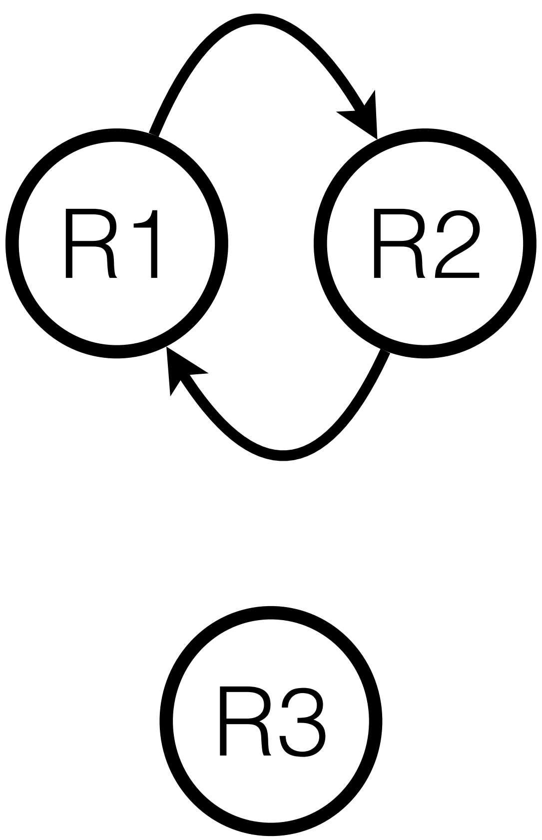
Rule Pruning Example

R1: $((\mathsf{name},\ \mathsf{name}),\ (\mathsf{officePhn},\ \mathsf{mobile},\ \mathsf{officePhn}),\ (=,=,=))$

R2: $((\mathsf{name},\ \mathsf{name}),\ (\mathsf{bornat},\ \perp,\ \mathsf{borncity}),\ (=,\not\approx,=))$

R3: $((\mathsf{nation},\ \mathsf{country}),\ (\mathsf{capital},\ \perp,\ \mathsf{capital}),\ (=,\not\approx,=))$

t3(Ian, ALT, Chine, Beijing, Hangzhou, 33668323)



iteration 1: {(R1, Yes), (R2, Yes), ~~(R3, No)~~}

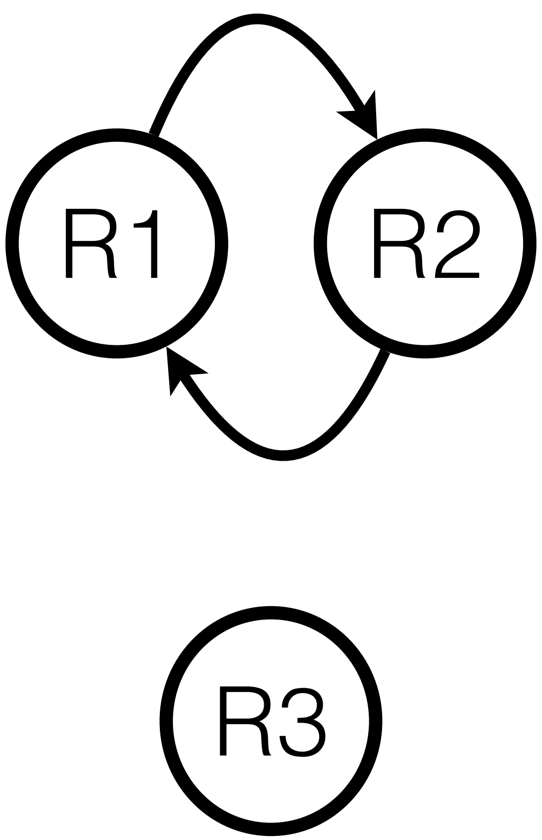
Rule Pruning Example

R1: ((name, name), (officePhn, mobile, officePhn), $(=, =, =)$)

R2: ((name, name), (bornat, $\perp$, borncity), $(=, \not\approx, =)$)

R3: ((nation, country), (capital, $\perp$, capital), $(=, \not\approx, =)$)

t3(Ian, ALT, Chine, Beijing, Hangzhou, 33668323)



iteration 1: {(R1, Yes), (R2, Yes),~~ (R3, No)~~}

iteration 2: {~~(R1, Yes),~~ (R2, No),~~ (R3, No)~~}

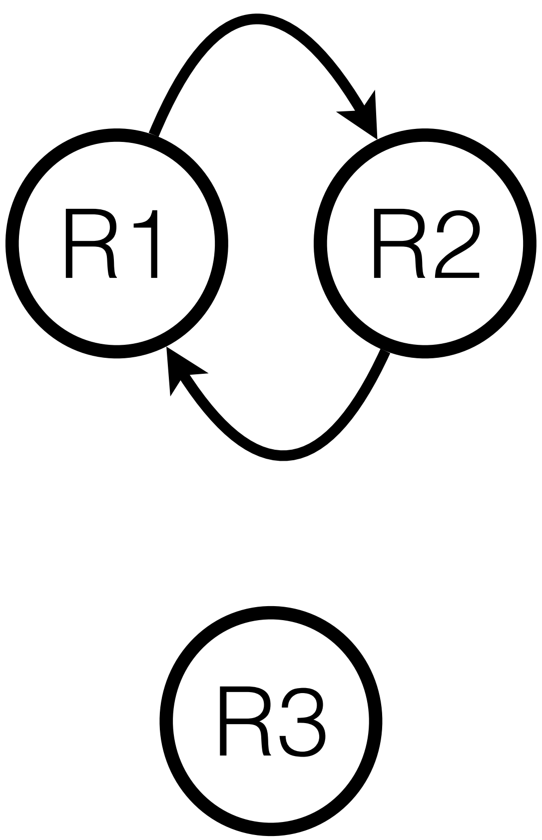
Rule Pruning Example

R1: $((\text{name}, \text{name}), (\text{officePhn}, \text{mobile}, \text{officePhn}), (=, =, =))$

R2: $((\text{name}, \text{name}), (\text{bornat}, \perp, \text{borncity}), (=, \not\approx, =))$

R3: $((\text{nation}, \text{country}), (\text{capital}, \perp, \text{capital}), (=, \not\approx, =))$

t3(Ian, ALT, Chine, Beijing, Hangzhou, 33668323)



iteration 1: {(R1, Yes), (R2, Yes), ~~(R3, No)~~}

iteration 2: {~~(R1, Yes),~~ (R2, No), ~~(R3, No)~~}

iteration 3: {~~(R1, Yes), (R2, No), (R3, No)~~}

Conclusion

- *Sherlock rules for accurately annotating and repairing data*
- *Fundamental problems*
- *Efficient algorithms*

Conclusion

- *Sherlock rules for accurately annotating and repairing data*
- *Fundamental problems*
- *Efficient algorithms*

Future Work

- *Let SQL drive the Sherlock workhorse*
- *Extend Sherlock rules to more data such as RDF (knowledge bases)*